# Homework 10.
# M7: TARTU-BUS-ANALYSIS

TEAM: Kairit Peekman, Kaspar Valk, Mikk-Kaspar Tammi

# Business understanding

## Business goals

**Background**
Our client is Tartu City Government, who launched new inner-city bus routes in 01.07.2019. The new bus route timetables were last changed 01.09.2019. Because the timetables are new and so little analysis has been done about them, then Tartu City Government can't be sure if these timetables are suitable to describe the actual bus movement.

**Business goals**
The goal of our project is to assess the quality of the bus route timetables and find out whether or not they need changes.

**Business success criteria**
Come to one of the following results and have reasoning to support the result:

1) The timetables do not need any changes as they are already nearly perfect and depict reality quite accurately.
2) The timetables need changes as there are one or several stops where the reality of the bus departure times is too often different from the timetables. In case of this conclusion, we will list all the problematic bus stops we found and provide convincing proof for each bus stop.

## Assessing our situation

**Inventory of resources**
We have two csv files about bus movement between 01.09.2019 and 18.10.2019 in Tartu. One of the files has data about all the occasions when a bus departed over 3 minutes later than supposed to and the other file has similar data but about the occasions when a bus departed over 10 seconds earlier than supposed to. The data includes info about the trip: number of the bus, date, id of trip and info about the stop: name and sequence of the stop, expected departure time, actual departure time and deviation. Our team consists of three students, all of who are taking a course Introduction to Data Science.

**Requirements, assumptions, and constraints**
Project deadline: submit a PDF-file of our results by Dec 16 at noon (12.00).
Project presentation time: present the results on Dec 19 at 14.00-17.00.

**Risks and contingencies**
If one of us loses access to their workplaces and data, then we have access to workplaces provided by UT and github to host data online.
If our project turns out to be too ambitious, then we will lower our goals.

**Terminology**
Bus line (or a route) - a sequence of bus stops that defines a single bus line.
Bus route trip - a single bus route completion, which is defined by the route number, start and end time (the same trip can occur multiple times over multiple days).

**Costs and benefits**
Not relevant as this is a student project.

# Data-mining goals

**Data-mining goals**
Produce 3 jupyter notebooks based on the following research questions:

1) Describe the situation of each singular bus route by trip information. Are the trips different (by the departure times from bus stops) by the day of the week and hour of the day?
2) Find out do any of the bus drivers or the physical buses themselves cause deviations in departure times from bus stops.
3) Describe the situation of each singular bus stop. Are there bus stops where the deviations of departure times are recurrently too large for multiple bus routes?

Based on the notebooks produce a singular poster that will have the results and conclusions.

**Data-mining success criteria**
The final result is concluded as a success if the reports provide answers with enough evidence to the questions they raised.

# Data understanding

## Gathering data

**Data requirements**

To fill our data-mining goals, we need data about all the occasions where the deviation of bus departure time was different from expected. The data should also include information about the bus driver, route number, trip number and departure times. As Tartu City Government changed the bus route timetables last in 01.09.2019 then the data needs to be from the according time period. The data should also be in a format that is easily readable (ie. as a csv file).

**Data availability**

We have access to the data required as Tartu City Government provided us the data files.

**Selection criteria**

Data that we will use comes in the form of two csv files that contain info about bus stops where the bus left too late or early. We plan on using fields: route_short_name, a_date, trip_id, trip_departure_time, stop_name, departure_time, estimated_departure_time, deviation, name, stop_sequence.
We will leave out id, route_long_name, stop_id, user_id, location_id, stop_code because we found no use of those.

## Describing data

The data is provided to us by Tartu City Government. The original source of the data is Ridango AS. The data is in two csv files.

The file having all the occasions when a bus departed too early has 125K rows.
The file containing all the occasions when a bus departed too late has 260K rows.
Both of the files have the same 16 columns and could be considered as a single csv file that has been split into two.

For our case we see the following columns to be of importance to us:

- **route_short_name**: number of the bus on that specific route.
- **a_date**: date when the trip happens.
- **trip_departure_time**: time when the bus leaves from the first stop and starts the trip.
- **stop_name**: name of the bus stop.
- **departure_time**: time when the bus is expected to leave that stop.
- **estimated_departure_time**: actual time when the bus left that stop.
- **deviation**: the difference between the time when the bus is supposed to leave the stop and the actual departure time.
- **name**: pseudonym of the bus driver.

- **trip_id**: ID of the trip that departs on specific time on a specific route regardless of date.
- **stop_sequence**: number of the bus stop that that bus is supposed to visit, when counting from the first stop.

## Exploring data

After combining the two csv files into a single larger file, the fields that are important to us have the following information:

**route_short_name**
Total 19 different route names as simple strings (ie, “7” or “9A”).

**a_date**
In total 47 different dates. The mean of occurrences on a single date is 8175, min is 4257, max is 10622.

**trip_departure_time**
In total 731 different trip departure times.

**stop_name**
In total 218 different stop names.

**departure_time**
In total 1299 different bus stop departure times. The mean of occurrences of single departure time is 302.

**estimated_departure_time**
In total 71859 different real departure time.

**deviation**
In total 125567 negative deviations (bus departed too early). In total there are 1506 different negative deviation times.
In total 266868 positive deviations (bus departed too late). In total there are 4267 different positive deviation times.

The column is saved as an object in format hh:mm:ss and to say more about the column, the data should be constructed as an integer in data preparation.

**name**
In total 161 different bus drivers. The mean of occurrences for a single name is 2437, min is 68 and max is 6198.

**trip_id**
In total 6157 different trip IDs. The max of occurrences for a single trip is 846, min is 1, mean is 64.

**stop_sequence**
In total 58 different numbers in stop_sequence.

## Verifying data quality

Considering the columns we chose to be of importance to us the overall quality of the data is quite good. The most problematic columns are **estimated_departure_time** and **deviation**. In esimated_departure_time there appear several such values which seem to have been contaminated: there are commas in seemingly random places. In deviation there are several values which are way too high or low to be logical: over an hour and so on. Also deviation is in unsuitable object format but it can be converted to a more suitable integer format in data preparation. There are no missing values for all the columns.

# Our plan

| Id | Task | Methods and tools | Mikk | Kaspar | Kairit |
|---|---|---|---|---|---|
| 1 | Data preparation and preprocessing. Remove unnecessary columns, fix problematic field values, convert data into suitable formats. | Python, Pandas, Numpy, Jupyter notebook | 2h | 2h | 2h |
| 2 | Make a Jupyter notebook based on the following research question:<br>Describe the situation of each singular bus route by trip information. Are the trips different (by the departure times from bus stops) by the day of the week and hour of the day? | Python, Pandas, Numpy, Matplotlib, Jupyter notebook | - | 20h | - |
| 3 | Make a Jupyter notebook based on the following research question:<br>Find out do any of the bus drivers or buses themselves cause deviations in departure times from bus stops. | Python, Pandas, Numpy, Matplotlib, Jupyter notebook | 20h | - | - |
| 4 | Make a Jupyter notebook based on the following research question:<br>Describe the situation of each singular bus stop. Are there bus stops where the deviations of departure times are recurrently too large? | Python, Pandas, Numpy, Matplotlib, Jupyter notebook | - | - | 20h |
| 5 | Conclude the found results and make the poster | Microsoft Word or Adobe Photoshop | 8h | 8h | 8h |